Research on advanced student writing across disciplines and levels

Introducing the *Michigan Corpus of Upper-level Student Papers*

Annelie Ädel & Ute Römer
Stockholm University / Georgia State University

This paper introduces the *Michigan Corpus of Upper-level Student Papers* (MICUSP) as a new resource that will enable researchers and teachers of English for Academic Purposes (EAP) to investigate the written discourse of highly advanced student writers whose written assignments have been awarded the grade 'A'. The usefulness of two aspects of the design of the corpus — variation across discipline and across student level — is illustrated by two case studies, one on attribution and one on recurrent phraseological patterns. The first case study investigates how references to the work of others are realized and to what extent disciplinary variation exists in unpublished academic writing by students. The second study examines the use of phraseological items (n-grams and phrase-frames) by students at four different levels of undergraduate and graduate study. The paper closes with a discussion of the results of both case studies and describes future avenues for MICUSP-based research.

Keywords: Michigan Corpus of Upper-level Student Papers (MICUSP), advanced student writing, English for Academic Purposes, attribution, phraseology

1. Introduction

This article introduces the *Michigan Corpus of Upper-level Student Papers* (MICUSP) as a new resource for research on advanced student academic writing. While unpublished academic writing produced by learners (as captured in ICLE, the *International Corpus of Learner English*) and published expert academic writing as covered in subsets of public-domain reference corpora such as the *British National*

International Journal of Corpus Linguistics 17:1 (2012), 3–34. DOI 10.1075/ijcl.17.1.01ade
ISSN 1384–6655 / E-ISSN 1569–9811

Corpus (BNC), the *Corpus of Contemporary American English* (COCA), or specialized corpora that are not normally publically available (e.g. the Hyland Corpus of research articles) have been given much attention in the fields of Corpus Linguistics and English for Academic Purposes, studies that investigate unpublished but advanced student writing are comparatively rare. Thanks to the impressive work carried out by researchers like Ken Hyland, John Swales and others, a great deal of information is available on academic genres such as the research article and the doctoral dissertation, but we know surprisingly little about the types of writing students produce on the way to becoming expert academic writers. Also, considering that "[f]ar more academic writing is produced for assessment purposes than for publication purposes" (Nesi et al. 2004: 440), it may seem striking that there are so few corpora available for researching assessed student writing (one notable exception here being the *British Academic Written English* corpus, BAWE; see Alsop & Nesi 2009).

One likely reason for the lack of research into advanced but unpublished student academic writing from different disciplines and student levels is the difficulty of accessing and systematically capturing this kind of writing. University instructors typically only have access to their own students' papers, i.e. assignments from a particular course in a particular discipline. In addition, it is not a trivial matter to collect and convert a large number of writing samples into a systematic and easily accessible collection (see Ebeling & Heuboeck 2007, Römer & O'Donnell 2011). MICUSP, the corpus we are focusing on in this article, addresses this problem and aims to at least partially fill the described gap in available resources.

We first describe the goals of the project and discuss the design and composition of MICUSP. We then illustrate the usefulness of two aspects of the corpus design — variation across discipline and across student level — by means of two case studies, one on attribution or references to other sources (conducted by the first author) and one on recurrent phraseological patterns (conducted by the second author). We close the paper by referring to ongoing research on MICUSP and by indicating some anticipated uses of the corpus.

2. MICUSP: A new corpus of advanced student academic writing

MICUSP is a collection of 829 A-graded student assignments totaling around 2.6 million words, representing different types of writing from a range of disciplines and student levels. The main rationale for the MICUSP project has been to make it possible to explore empirically and on a broader scale the question when it is that "students really begin to write like academics" (John Swales 2009, personal communication).[1] It was conceived of as "an inductive exercise […] building a picture

of disciplinary variation (or not) and of increasing demands through rank (or not) as samples are collected and analysed" (John Swales 2004, personal communication). We will return to this in Section 4 when summarizing the two case studies, as they specifically deal with variation across discipline and student rank or level, and thus directly explore the core topics for which MICUSP was intended.

The goal of the project has been to collect a large and relatively balanced selection of samples of academic writing in English from different disciplines by senior undergraduates and by students at different stages of graduate level study. The context is that of a major research university in the United States: the University of Michigan in Ann Arbor. The university's English Language Institute is the host and sponsor of the corpus.[2]

The corpus is meant to represent "proficient" writing in that only assessed A-grade papers have been accepted for inclusion. Even though there are many different reasons why a given paper may be awarded an 'A' — not all of which necessarily have to do with writing skills per se; reasons may range from grade inflation to the use of thought-provoking ideas — the assumption was that the A-grade samples should demonstrate the standards set by instructors across different departments at the University of Michigan and constitute successful pieces of writing (i.e. assignments with which a student passes a class). It was also considered important that submitted texts not be too heavily edited by a teacher or supervisor, as may be the case in the final version of a Master's thesis, for example. This was checked by asking in the participant questionnaire whether any input from others — and, if so, by whom and how much — was received in the process of writing.

The fact that MICUSP represents a group of "EAP corpora [which] contain only student texts that have been awarded high grades" (Krishnamurthy & Kosem 2007:366) has been criticized: "[…] without lower-grade student texts, there is no opportunity for monitoring progression, or for making comparisons with the higher-grade student writing. And after all, the students receiving lower grades are precisely the ones that require more EAP input/help, and that we should be more concerned with" (ibid.: 367). We believe that both types of assessed student writing — low-scoring as well as high-scoring — need to be represented in corpora, thus enabling empirical research of different kinds. Somewhat simplified, we can call the two types of data "problem writing" and "target writing". Corpora of the MICUSP type are of great potential value to EAP teachers around the world in that they can be seen as representing target writing. They are also useful collections of writing samples for less proficient or lower-level academic writers who may be looking for model papers of a certain type in their field of study (e.g. lab reports in Engineering) or need to know how a particular word or phrase is used appropriately in an academic paper.

MICUSP is not balanced for genre, but any type of assessed writing that students have submitted has been accepted. That way, we were able to capture which types of assessed writing students across disciplines and levels are expected to produce on their way to graduation. In a process of paper classification described in Römer & O'Donnell (2011), all papers were annotated with a list of textual features and assigned a paper type label. The following seven different paper types were found to be represented in the corpus: argumentative essay, creative writing, critique/evaluation, report, research paper, research proposal, and response paper.[3]

It was an important goal in the compilation process to achieve a relatively balanced distribution with respect to discipline. This turned out to be quite difficult, however, one reason being that there is considerable variation in both the number and the length of assessed papers produced across disciplines. For example, in Physics, very little assessed writing is done at all and it is mainly in the form of short papers (2,145 words on average), most of which are reports. In History, by contrast, much longer papers are produced with an average of 4,566 words in our dataset. It is clear that genre output is very much dependent on shifting values across disciplines.[4] Another reason why achieving an exact balance across disciplines was not feasible was that the number of students enrolled in academic programs differs considerably from department to department. In the end, we were able to collect between 21 (Physics) and 104 (Psychology) papers from each discipline.

The papers were collected from sixteen disciplines across the four academic divisions in place at the University of Michigan: Humanities and Arts, Social Sciences, Biological and Health Sciences, and Physical Sciences. To achieve an acceptable representation of the university's wide disciplinary spectrum and to capture a range of different types of writing, we selected between four and five disciplines from each division.[5] We believe that the selected disciplines provide an adequate representation of what a comprehensive and diverse institution the University of Michigan is. Table 1 lists the sixteen disciplines and shows the number of papers and words for each discipline and for the four academic divisions (see also Römer & O'Donnell 2011). The overall word count is highest for the Social Sciences division (978,254), followed by the Humanities and Arts (734,437) and Biological and Health Sciences (511,550), and lowest for the Physical Sciences (392,288). The strongest disciplines in MICUSP, in terms of numbers of papers, are Psychology (104 papers), English (98), Sociology (72), and Biology (67). The smallest disciplinary subsets of papers come from Physics (21), Economics (25), Civil and Environmental Engineering (31), and Mechanical Engineering (32). For the remaining disciplines, the paper counts range from 40 (History and Classical Studies) to 62 (Natural Resources and Environment; Political Science).

Table 1. The composition of MICUSP: Academic divisions and disciplines

Discipline	Papers		Words	
Humanities and Arts				
English (ENG)	98	11.8%	268,733	10.3%
History and Classical Studies (HIS)	40	4.8%	182,629	7.0%
Linguistics (LIN)	41	4.9%	155,047	5.9%
Philosophy (PHI)	44	5.3%	128,028	4.9%
Total	223	26.8%	734,437	28.1%
Social Sciences				
Economics (ECO)	25	3.0%	78,070	3.0%
Education (EDU)	46	5.5%	150,282	5.7%
Political Science (POL)	62	7.5%	210,783	8.1%
Psychology (PSY)	104	12.5%	323,326	12.4%
Sociology (SOC)	72	8.7%	215,793	8.2%
Total	309	37.2%	978,254	37.4%
Biological and Health Sciences				
Biology (BIO)	67	8.1%	176,124	6.7%
Natural Resources and Environment (NRE)	62	7.5%	176,653	6.8%
Nursing (NUR)	42	5.1%	158,773	6.1%
Total	171	20.7%	511,550	19.6%
Physical Sciences				
Civil and Environmental Engineering (CEE)	31	3.7%	98,918	3.8%
Industrial and Operations Engineering (IOE)	42	5.1%	124,973	4.8%
Mechanical Engineering (MEC)	32	3.9%	123,335	4.7%
Physics (PHY)	21	2.5%	45,062	1.7%
Total	126	15.2%	392,288	15.0%
All	829		2,616,529	

MICUSP papers were collected from student writers at four different levels of study: senior undergraduate students (i.e. pre-graduate level students in their fourth and final year of study), first-year, second-year, and third-year graduate students. Table 2 shows the number of papers and words per student level. There is an almost even division into 432 (52.1%) undergraduate and 397 (47.9%) graduate student papers. Among the three graduate levels, figures decrease from first to third year, with 203 papers (24.5%) written by first-year, 117 (14.1%) written by second-year, and 77 (9.3%) written by third-year graduate students. The distribution of papers across disciplines is somewhat different for each level of study.

Compared to the overall distribution (see Table 1), there are relatively more papers from Biology, English, Philosophy, and Political Science in the set of final-year undergraduate submissions. Among the first-year graduate student submissions, numbers are highest for Natural Resources and Environment and for Sociology. Second-year graduate student papers are somewhat overrepresented in the set of Psychology papers and extremely rare in the sets of Civil and Environmental Engineering and Philosophy papers. The highest share of third-year graduate student papers can be found among the Sociology submissions.

Table 2. The composition of MICUSP: Four levels of study

Student level	Papers		Words	
Final-year undergraduate (G0)	432	52.1%	1,063,354	40.6%
First-year graduate (G1)	203	24.5%	747,747	28.6%
Second-year graduate (G2)	117	14.1%	446,336	17.1%
Third-year graduate (G3)	77	9.3%	359,092	13.7%
All	829		2,616,529	

Like many US universities, the University of Michigan enrolls a relatively large number of international students each year; 12.7% of the student body counts as international according to figures from Fall 2009.[6] This circumstance is reflected in the corpus, as it includes not only English native-speaker texts, but also non-native speaker contributions. The average proportion of non-native speaker contributions in the corpus is 17.9%, with higher percentages at graduate level (25.6% at G2; 28.6% at G1 and G3). The native-speaker status of the student writers has been recorded and can be used as a variable in research.

The corpus files have been XML-encoded and annotated for variables such as discipline, level of study, nativeness and text type. Another feature of the corpus which deserves special mention is the mark-up of all quoted material. We have considered it important to be able to keep separate, on the one hand, stretches of text produced by the current authors (i.e. the student writers represented in the corpus) and, on the other, stretches of text produced by any other authors that the current author has chosen to cite (this could also involve spoken utterances made by an informant or written replies to a questionnaire, for example in Sociology or Linguistics papers). The marked-up quoted material enables searches on the student writers' own words exclusively, and not including quotes from expert academics, informants, or poets (for further discussion and examples, see Ädel 2010b).

Like the *Michigan Corpus of Academic Spoken English* (MICASE; Simpson et al. 1999), MICUSP is freely available to the research community through a user-friendly search and browse interface.[7] Several versions of the corpus will also soon

be distributed for offline use on a CD-ROM (Römer & O'Donnell, in preparation). For further technical specifications of the corpus, see O'Donnell & Römer (Forthcoming), and Römer & O'Donnell (2011).

3. Two case studies illustrating the variables "discipline" and "student level"

In this section, we present two case studies based on MICUSP. The first case study (Section 3.1) deals with variation across disciplines, specifically with respect to the use of attribution, or references to other sources. The second case study (Section 3.2) deals with variation across student ranks, looking at phraseological developments as students move from final-year undergraduate to third-year graduate level. The two case studies are not meant to be connected thematically, but illustrate two different aspects of the design of the corpus.

3.1 Attribution and disciplinary variation

Recent research has shown that 'academic discourse' should not be seen as a monolith, but that the academic discipline exerts quite some influence over discourse conventions. The academic discipline is seen as a source of variation which can be even stronger than that of first language and cultural background (see, e.g., Dahl 2004). One of the features in which considerable disciplinary differences have been found is the use of attribution, or references to other sources, in published academic writing (e.g. Hyland 1999).

Attribution involves "reference to prior research", specifically "the attribution of propositional content to another source" (Hyland 1999: 341). There are many different ways in which propositional content can be attributed to another source, illustrated in the following examples from MICUSP:

(1) a. Marx argued that general human emancipation is contained within the emancipation of society from private property ("slavery"), which would take its political form in the emancipation of the workers (Marx 1977c: 85). [Sociology]

b. Bearman and Bruckner (2002) state that adolescent male opposite-sex twins are twice as likely to report same-sex attraction compared to similar-sex twins. [Psychology]

(2) a. He suggests that meaning in life is, "generally discovered in creative pursuits, in life's experiences and relationships, and in attitudes taken toward both positive life experiences and the "tragic triad" of pain/suffering, guilt, and death." [Sociology]

b. They claim that their harvesting circuit improves direct charging of a storage medium by nearly 400%. [Mechanical Engineering]

(3) a. This bacterium is a non-motile, non-sporeforming, Gram-negative coccobacillus (Bahmanyar and Cavanaugh, 1976). [Biology]
b. Some 526 Western European pilgrimage accounts written between 1100 and 1500 survive, and the vast majority concern the Jerusalem pilgrimage.3 [History][8]
c. The velocity distribution in the atomic beam will only result in a rescaling of the time axis, which does not affect the mixing angle [3]. [Physics][9]

The distinction between 'integral' versus 'non-integral' types of citation forms (Swales 1990) is a formal one, referring to types that are syntactically integrated into the text (examples in 1 and 2 above) versus types that are syntactically non-integrated (examples in 3 above). The latter is typically separated by parentheses, or indexed by footnotes or numbers in brackets.

Being such a central feature of academic writing, attribution has been approached from different perspectives, including both theoretical ones relating to the social construction of knowledge and highly applied ones relating to how to teach it to novice academics. In fact, learning how to make appropriate citations is part of learning the ways in which knowledge is constructed and communicated in one's discipline (cf. Hyland 1999). Whether the MICUSP students seem to have learnt this important academic skill is a question that will be explored here by means of corpus-linguistic techniques. As we have seen, the MICUSP students are quite far advanced in their academic study, but they are not professional academics writing for publication.

The main research question of this study is: To what extent does disciplinary variation in the use of attribution exist in unpublished academic writing by students? Put differently, we want to find out whether the MICUSP students' writing exhibits similar variation in attribution across disciplinary boundaries as that of academic professionals. This question does not only apply to the *frequency* of attribution, but also to the linguistic *form* used. Previous research (e.g. Hyland 1999) has found that, in published academic writing, disciplinary differences also obtain with respect to the realization of attribution. This leads to the second research question: How are references to the work of others realized in student writing?

3.1.1 *Method and material*

The method involved concordancing and extracting instances of attribution in the form of (a) third person pronouns, (b) proper names (including *ibid**), (c) a selection of nouns potentially used in attribution (e.g. *author*; researcher*; investigator**), and

(d) numerical citation forms (*19**; *200**). The proper names were retrieved from the list of references or from footnotes/numbered brackets of the papers included, in combination with manual selection in papers where such lists were not included. Only one instance per attribution was counted, although using (a) through (d) above often captured the same example more than once (cf. the running text of 1a, 1b, and 3a above); for instance, should both a proper name and the year of publication be involved in attributing content to a source. References to academic texts per se, as in *book, work, article*, or even *it*, were not included in the study, as they were deemed not likely to retrieve relevant examples of attribution. When *article* occurs, for example, it typically does so in connection with author name (*Hanselman's article cites four studies…*) or footnote/bracketed number (*The article [8] continues…*), or the reference is metadiscursive, i.e. refers to the current text, as in *This article reviews the effects that maternal nutrition can produce in the fetus.*

Each retrieved token was checked to make sure that it did indeed function as attribution. Considering that a great deal of manual analysis is required for a study of attribution, it was not feasible to examine the entire corpus. Instead, samples of papers (total N=208) taken from ten different disciplines were analyzed, consisting of approximately 680,000 words, which amounts to 26% of the entire corpus. Table 3 shows the distribution of papers analysed across the disciplines.

Table 3. Distribution of papers used for the attribution study across the ten disciplines, including the number of words per discipline (abbreviations are glossed in Table 1)

Academic division	Discipline	Papers	Words
Humanities & Arts	LIN	31	110,310
	PHI	20	47,138
	HIS	18	95,860
Social Sciences	SOC	31	91,512
	PSY	28	96,141
Physical Sciences	IOE	21	77,608
	CEE	13	39,759
	PHY	11	19,791
Biological & Health Sciences	BIO	25	56,836
	NUR	10	44,108

The samples were selected in an opportunistic manner in that they were the first papers to be collected; at a certain point during the creation of the corpus, these amounted to all of the papers represented in these disciplines. No discipline is represented by fewer than ten papers. All discipline samples have at least 40,000 words, most of them considerably more than that, with the exception of Physics,

where not only is very little writing done, but texts also tend to be quite short (cf. Section 2 above).

3.1.2 *Results*

Concerning the first research question, how frequently attribution is used across disciplines, Figure 1 shows the normalised amount of attribution across disciplines.

As we can see, there is a considerable frequency range from as few as 53 in Industrial & Operations Engineering to as many as 323 occurrences in History.[10] Attribution is clearly a feature that gives rise to disciplinary variation, at least in quantitative terms. Somewhat arbitrarily, we can talk about three different magnitudes based on these numbers: one in the under 100 range (Industrial & Operations Engineering; Physics; Linguistics; and Civil & Environmental Engineering), one in the 100–200 range (Nursing; Psychology; Biology; Sociology; and Philosophy), and one at over 300 occurrences (History).

In looking for further patterns, it is interesting to consider whether the distribution mirrors academic divisions. Taking this view, the Humanities disciplines occur at the top end, with the exception of Linguistics. The Social Sciences disciplines cluster after the Humanities. The two disciplines representing the Biological and Health Sciences occur next, and they are followed by the three disciplines in the Physical Sciences and Engineering, which occur at the bottom end. Thus, there is a clear frequency correlation between attribution and broad academic divisions; the linguistic patterning largely follows the general categorisation of disciplines in the academic divisions. The only "outlier" in this respect is Linguistics.

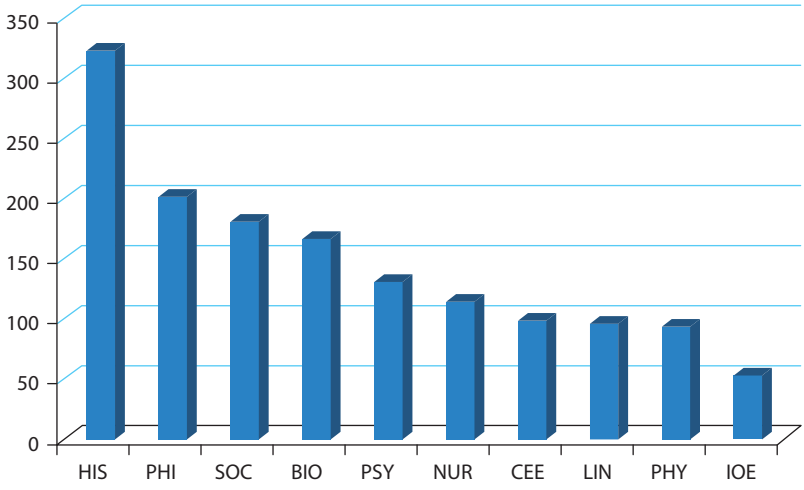

Figure 1. Number of attribution units across disciplines (F per 10,000 words)

If we turn next to the second research question of how attribution is realized, Figure 2 shows the results for integral and non-integral types across disciplines (sorted by descending frequency of integral forms).

Just like the overall number of attribution units, the integral and non-integral types display considerable disciplinary variation. In six of the ten disciplines, the integral type is clearly the unmarked form. There is a clear cut-off point between Psychology and Nursing. In fact, Psychology is an interesting borderline case, where both types are almost equally common.[11] In three of the disciplines — Nursing, Biology and Civil & Environmental Engineering — the non-integral type clearly predominates.

If we look at the academic divisions from the perspective of integral/non-integral types, a scattered picture emerges. The clear patterning found for the overall occurrence of attribution is not reproduced here. There is, perhaps, one exception in that Nursing and Biology (which are both in the Biological & Health Sciences) have quite similar distributions, but apart from that the variation found does not correlate with academic division.

It is interesting to consider whether the disciplinary patterns found in this type of advanced student writing are also evident in published academic writing, even though we must keep in mind that there may be considerable genre differences between these two general types of writing. With slight reservation concerning possible differences in the definition of attribution, such a comparison can be made based on data presented in Hyland (1999: 347). Hyland's corpus of research articles can be compared to MICUSP with respect to six disciplines: Philosophy, Sociology, Linguistics, Biology, Physics and Engineering.[12]

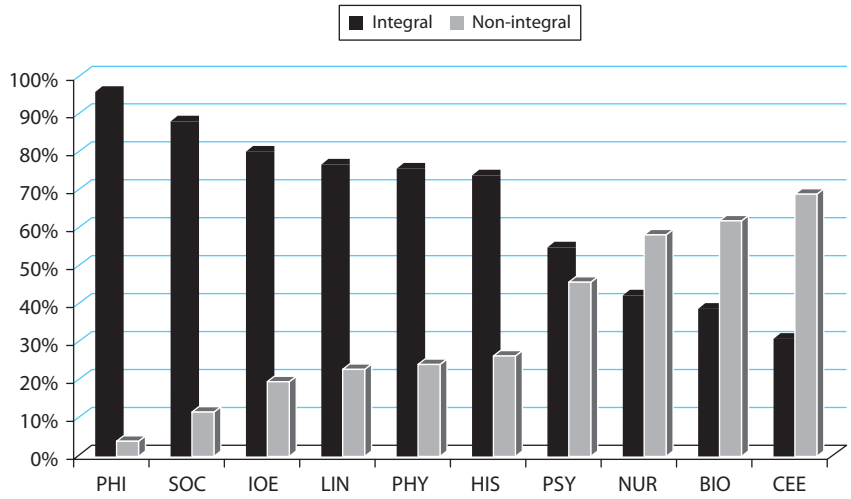

Figure 2. Integral vs. non-integral types of attribution across disciplines (percentages)

Figure 3 shows the normalized numbers of attribution units across the six compared disciplines in MICUSP and the Hyland corpus.

Overall, the frequencies in the two corpora are quite comparable. Four of the six disciplines are particularly closely matched: Biology, Linguistics, Physics and Engineering. Philosophy and Sociology, however, show greater discrepancies between the professional and the advanced student writing. It could be that attribution is such a salient feature of these fields, thus making it easier for students to overdo it when trying to model "philosophy writing" and "sociology writing" (cf. Ädel & Garretson 2006) — recall that Figure 2 showed that Philosophy and Sociology have the largest proportions of integral types of all ten disciplines. Part of the explanation may also have to do with disciplinary preferences for citations of recent work versus foundational documents (Hargens 2000). It would seem that, in both Philosophy and Sociology, the foundational documents enjoy high status; exegesis is a prominent activity and questions raised in classic texts are often pursued (cf. Hargens 2000: 847). Discrepancies could also be caused by different conventions being enforced by professors and departments (in the case of MICUSP data) or by editors/publishers (in the case of the Hyland data), for example leading to greater use of non-integral forms.

It is important to make the point that it would not make sense to speak of student "overuse" in Philosophy and Sociology, without first doing some manual analysis to establish whether the use of attribution is inappropriate or infelicitous. Variables other than discipline may be at work. Consider genre. For example, the student texts in Philosophy may be required to interact with other sources to a great extent — say, for a literature review — so it may be part of the purpose of a given text to draw on attribution heavily. To illustrate how other variables could be

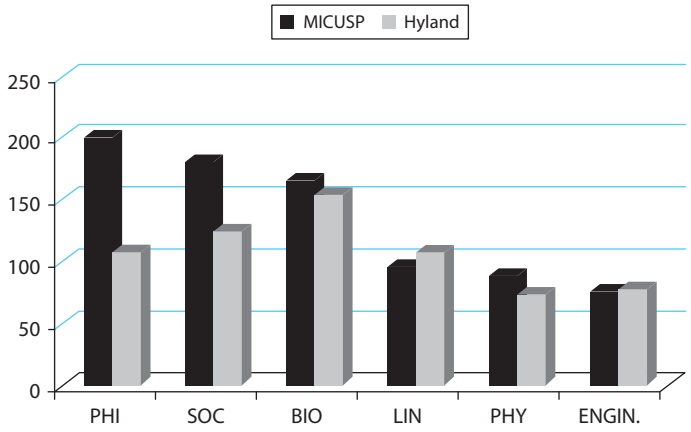

Figure 3. Number of attribution units across disciplines (F per 10,000 words) in the two corpora

explored in MICUSP, we can make use of the online search tool MICUSP Simple. In the case of the Philosophy essays, a consideration of the general "paper type" classifications does not seem to yield any insights. However, if we consider the Philosophy titles, it becomes clear that the "major thinkers" are quite prevalent. Some of the titles, across paper types, include the following: *Hume and Smith on Justice; Two Distinctions in Kant; Grice's Analysis of Metaphor and Irony; George Schlesinger and Pascal's Wager; Aristotle on Friendship; Examining Socrates' Desire Theory;* and *On Hempel's Account of Scientific Explanations*. The emphasis on classic texts/thinkers is clearly there, and one would expect such an emphasis to be even stronger in an educational context (MICUSP) compared to a research context (Hyland corpus).

If we look at Linguistics, for comparison, the paper type distribution shows that there is a relatively large proportion of research papers (37%) in this group, which is unlike the Philosophy group where no research papers at all are represented. Presumably, in a research paper, a large chunk of the text will have to be dedicated to the study in question and its results, thus leaving less space for references to previous work by others compared to a pure literature review. Furthermore, the Linguistics titles reveal no instances of references to experts in the field — again, by contrast to Philosophy — with the exception of three critiques (representing a mere 7% of the Linguistics subcorpus). A sample of Linguistics titles include *Writing About Portrait Paintings: A Discourse Analysis; Conversation as a 'Collaborative Achievement'; A Discussion of the VP-External and VP-Internal Subjects Hypotheses,* and *German Fricatives*.

Thanks to relatively rich documentation, MICUSP Simple can be used to examine various variables that potentially influence the use of language. There are many resources for doing this through the online search interface. In addition to discipline and student level, which each paper has been annotated for, genre can be approached by "paper type", possibly also "textual features", and also complemented by checks of titles and, needless to say, manual analysis of the texts themselves.

3.1.3 *Summary*

Considerable differences were found in how often attribution occurred across disciplines in MICUSP. This is a pattern that has been reported for published academic writing (Hyland 1999), and the MICUSP data show that it also holds true for unpublished academic writing at this highly proficient level. Considerable differences in the realization of attribution, whether integral or non-integral types were used, were also found on the basis of discipline. Comparing part of the results to data from Hyland (1999) on published academic writing, essentially the same patterns were found in proficient student writing at least in four out of six disciplines.

If we recall the quote from Swales offered at the beginning of this paper, we can note that the MICUSP project was characterized as enabling the painting of a "picture of disciplinary variation (or not) […] as samples are collected and analysed". With a subset of MICUSP examined, it becomes clear that the parenthetical "or not" can be scratched, as quite extensive disciplinary variation has been found in terms of both overall frequency and formal realization. We can add that some of the variation is likely due to genre differences, as indicated in the discussion of the Philosophy and Linguistics subsets. The question when it is that students really begin to write like academics can, to some extent, be answered here, in that MICUSP student writers (taken as a conglomerate) already appear to be highly specialized and well socialized into their respective disciplines — at least from the perspective of attribution. Other features, however, may be adopted and mastered at quite different stages.

Previous studies of attribution in unpublished student writing have focused on problematic uses of sources, such as paraphrasing difficulties (Campbell 1990) or plagiarism (Pecorari 2006), but our findings offer a less problem-oriented perspective in that the MICUSP students actually appear largely to conform to conventions similar to those used by professionals.[13] One could see this as students at this level getting the overall pattern right, although there may still be specific problems to be dealt with, for example in terms of paraphrasing. Of course, the MICUSP sample represents papers which have been awarded the highest grade; the story would likely have been different had we also examined a set of low-scoring papers.

3.2 Phraseological items and student level

Our second case study examines the distribution of frequent phraseological items across sets of MICUSP papers produced by students at different levels of undergraduate and graduate study. We use the notion "phraseological item" to refer to recurring contiguous or non-contiguous combinations of two or more words. The version of MICUSP used in this case study consists of four subcorpora, one each of papers written by final-year undergraduate students, first-year graduate students, second-year graduate students, and third-year graduate students (see Table 2 above for details). All analyses were carried out separately for each of the four sets of papers and results were then compared across sets. We also submitted the Hyland corpus of published research articles (Hyland 1998) to the same kind of analysis as the MICUSP subcorpora. The Hyland corpus (henceforth Hyland) contains 240 research articles from eight academic disciplines (Biology, Electrical Engineering, Linguistics, Marketing, Mechanical Engineering, Philosophy, Physics, Sociology) and has a size of just over 1.3 million words. The resulting figures

derived from expert rather than novice academic writing will serve as a reference in our data comparisons.

We started out from the assumption that the main meaning-carrying unit in language is not the word in isolation but the phrase, i.e. a sequence of two or more words which may allow for internal variation (e.g. *it would be interesting*, *it would be very interesting*). Like Sinclair (2008: 408), we consider the phrase "central and pivotal" in language description, and acknowledge the powerfulness of an approach that focuses on phrases or clusters, rather than isolated words. We thus add to the growing body of research that is dedicated to uncovering the extent of the "phraseological tendency" of language (Sinclair 1996), and points to an inseparability of lexis and grammar (see, e.g., Biber et al. 2004, Hoey 2005, Hunston & Francis 2000, Granger & Meunier 2008, Römer 2009b and 2010, Römer & Schulze 2009, Sinclair 1991 and 2004, Stubbs 2001).

Phraseological items also have an important status in pedagogical contexts. As Hyland (2008: 41), among others, has pointed out, "[c]lusters seem to present considerable challenges to student writers struggling to make their texts both fluent and assured to readers in their new communities". At the same time, it is important for novice academic writers to know which phraseological items are used by more advanced and successful writers and which are accepted by expert readers (initially their university instructors and examiners). It would hence appear essential, from a pedagogical perspective, to know what the most common phraseological items in proficient student writing are and how they change from undergraduate to graduate levels.

In our analysis of phraseological items we included n-grams, i.e. sequences of n words (e.g. *as well as*, *on the other hand*), and phrase frames (short p-frames), i.e. sets of n-grams which are identical except for one word in the same position (e.g. *on the one hand* and *on the other hand* are part of the p-frame *on the * hand*), of different lengths. For a selection of high-frequency p-frames, we studied the lists of p-frame variants, i.e. the lexical items that fill the * slot in the frame. This part of the analysis provided insights into pattern variability and helped us see to what extent Sinclair's Idiom Principle (Sinclair 1987, 1991, 1996) is at work, i.e. how fixed language units are or how much they allow for variation. The software tool we used to extract lists of n-grams and p-frames (with their respective variants) from the MICUSP subsets was *kfNgram* (Fletcher 2002–2007). Our definition of p-frame, however, differs from Fletcher's in that we only consider n-grams with an *internal* variable slot (i.e. A*CD and AB*D for the 4-gram ABCD) to be p-frames (following Römer 2010), which made some post-processing of the *kfNgram* output lists necessary.[14] We extracted frequency lists of n-grams and p-frames of spans 3, 4, and 5. Since the 5-gram and 5-p-frame lists were comparatively short and mainly consisted of topic-related items, such as *asymmetry in stock price response* or *the*

** of domestic violence*, we will only report on findings based on 3-/4-grams (e.g. *in terms of*, *as a result of*) and 3-/4-p-frames (e.g. *in * of*, *as a * of*). In this study, topic-related items are not of prime interest because we are concerned with potential differences in discourse preferences in student writing at different levels of study.

3.2.1 *N-grams across student levels*

The first part of the analysis consisted of a close examination and comparison of ranked lists of frequent 3- and 4-grams in four MICUSP subsets containing papers written by students at final-year undergraduate (G0) and first (G1), second (G2) and third (G3) year graduate levels. In Tables 4 and 5 presented below, n-grams are listed in order of frequency of occurrence in each of the MICUSP subsets. For each n-gram we checked how well dispersed it was across a MICUSP subset, and only included items in our lists that occurred in at least five different texts. This dispersion factor filtered out n-grams which appeared in the *kfNgram* output lists only because they were highly frequent in one or two student papers. These n-grams were hardly typical of student writing at a particular level of study but rather represented writer idiosyncrasies or topic-related phrases and technical terms. To give just two examples, the 3-gram *part time faculty* originally ranked second in the 3rd year graduate 3-gram list but was used 142 times by one student and was hence taken off the list. The same applied to *stock price response*, a high-frequency item in the 2nd year graduate 3-gram list for which 79 occurrences were found in one single paper. The same procedure was applied to the corpus of published academic writing (Hyland). Lists of the 20 most frequent 3- and 4-grams in Hyland are displayed in Table 6 together with normalized frequencies.

Table 4. Top-20 3-grams across MICUSP student levels

Rank	Final year UG	Norm freq*	1st year graduate	Norm freq*	2nd year graduate	Norm freq*	3rd year graduate	Norm freq*
1	IN ORDER TO	61.7	IN ORDER TO	43.2	AS WELL AS	39.4	IN ORDER TO	40.3
2	AS WELL AS	40.7	AS WELL AS	41.2	IN ORDER TO	37.2	THE UNITED STATES	30.5
3	THE UNITED STATES	34.9	ONE OF THE	25.4	IN TERMS OF	23.9	*it is not*	26.0
4	ONE OF THE	27.6	THE UNITED STATES	25.4	*more likely to*	21.4	AS WELL AS	23.8
5	*the number of*	27.6	DUE TO THE	22.5	ONE OF THE	21.4	ONE OF THE	23.2
6	*the fact that*	25.5	IN TERMS OF	21.3	*the fact that*	18.2	*on the other*	21.3
7	*there is a*	18.4	*based on the*	20.3	THE UNITED STATES	18.0	*that it is*	20.7
8	DUE TO THE	18.1	*the fact that*	19.1	*there is a*	17.6	*of the state*	19.3
9	IN TERMS OF	17.3	*there is a*	18.0	*it is not*	16.9	*the idea of*	19.3

Table 4. (*continued*)

Rank	Final year UG	Norm freq*	1st year graduate	Norm freq*	2nd year graduate	Norm freq*	3rd year graduate	Norm freq*
10	THE USE OF	17.1	*part of the*	17.8	*the number of*	16.4	*the other hand*	19.3
11	*be able to*	17.0	*in the united*	17.5	*on the other*	15.3	IN TERMS OF	19.0
12	*all of the*	15.8	THERE IS NO	16.7	THERE IS NO	14.9	*the effect of*	16.8
13	*as a result*	15.6	*as a result*	15.9	*the development of*	14.6	*based on the*	16.2
14	*in the united*	15.1	*the number of*	15.8	*part of the*	14.4	*the role of*	16.0
15	*part of the*	14.9	*the importance of*	15.2	*based on the*	13.5	THE USE OF	16.0
16	*to be a*	14.8	*the role of*	15.2	THE USE OF	13.5	*the relationship between*	15.1
17	*it would be*	13.9	THE USE OF	14.9	*the effects of*	13.3	THERE IS NO	15.1
18	*that it is*	13.9	*be able to*	14.7	*the other hand*	13.1	*in other words*	14.3
19	*some of the*	13.4	*it is not*	14.1	DUE TO THE	13.1	*the case of*	14.3
20	THERE IS NO	13.4	*the relationship between*	13.6	*be able to*	12.8	DUE TO THE	14.0

* In order to facilitate comparison, all n-gram counts have been normalized to counts per 100,000 words.

Table 5. Top-20 4-grams across MICUSP student levels

Rank	Final year UG	Norm freq*	1st year graduate	Norm freq*	2nd year graduate	Norm freq*	3rd year graduate	Norm freq*
1	IN THE UNITED STATES	14.5	IN THE UNITED STATES	17.0	ON THE OTHER HAND	12.8	ON THE OTHER HAND	19.3
2	*as well as the*	11.8	ON THE OTHER HAND	10.8	AT THE SAME TIME	10.1	IN THE UNITED STATES	12.9
3	ON THE OTHER HAND	9.4	IN THE CASE OF	9.0	IN THE UNITED STATES	9.5	IN THE CASE OF	10.9
4	IT IS IMPORTANT TO	8.8	*as well as the*	8.5	*as well as the*	7.9	AT THE SAME TIME	9.5
5	*the end of the*	8.6	AT THE SAME TIME	8.1	*are more likely to*	6.3	*in the context of*	8.1
6	*the university of michigan*	7.8	AS A RESULT OF	6.6	*the end of the*	6.3	IT IS IMPORTANT TO	7.3
7	AS A RESULT OF	7.7	*in the form of*	6.2	IN THE CASE OF	5.4	AS A RESULT OF	6.7
8	AT THE SAME TIME	7.5	*the end of the*	6.1	*more likely to be*	5.2	*of the united states*	6.7
9	IN THE CASE OF	7.1	*the university of michigan*	5.8	*in the context of*	4.7	*i would like to*	6.2
10	AT THE END OF	6.3	IT IS IMPORTANT TO	5.8	IT IS IMPORTANT TO	4.7	*for the purposes of*	5.0

Table 5. (*continued*)

Rank	Final year UG	Norm freq*	1st year graduate	Norm freq*	2nd year graduate	Norm freq*	3rd year graduate	Norm freq*
11	*the beginning of the*	5.7	*AT THE END OF*	5.1	*AT THE END OF*	4.3	*the role of the*	5.0
12	*one of the most*	5.3	*are more likely to*	4.8	*in terms of the*	4.3	*as a function of*	4.8
13	*the rest of the*	5.1	*it is clear that*	4.4	*the university of michigan*	4.3	*on the one hand*	4.8
14	*it is clear that*	4.8	*one of the most*	4.4	*AS A RESULT OF*	4.1	*AT THE END OF*	4.5
15	*at the university of*	4.6	*the nature of the*	3.9	*the fact that the*	4.1	*can be used to*	4.5
16	*to the fact that*	4.6	*to the fact that*	3.9	*will be able to*	4.1	*in terms of the*	4.5
17	*can be seen in*	4.4	*will be able to*	3.9	*at the university of*	3.8	*with respect to the*	4.5
18	*that there is a*	4.4	*at the university of*	3.8	*is more likely to*	3.8	*in addition to the*	4.2
19	*the fact that the*	4.3	*can be used to*	3.8	*were more likely to*	3.8	*in the process of*	4.2
20	*to be able to*	4.3	*the fact that the*	3.8	*the rest of the*	3.6	*in a way that*	3.9

* In order to facilitate comparison, all n-gram counts have been normalized to counts per 100,000 words.

Table 6. Top-20 3-grams and 4-grams in Hyland

Rank	3-gram	Norm freq*	4-gram	Norm freq*
1	*as well as*	28.6	*in the case of*	15.8
2	*the number of*	27.8	*on the other hand*	14.0
3	*in order to*	27.5	*at the same time*	10.4
4	*in terms of*	26.8	*on the basis of*	8.5
5	*the use of*	25.5	*as a function of*	7.1
6	*one of the*	23.0	*in the united states*	7.1
7	*the fact that*	22.2	*in terms of the*	6.9
8	*shown in fig*	20.7	*the extent to which*	6.7
9	*the case of*	20.2	*as shown in fig*	6.0
10	*there is a*	19.4	*with respect to the*	6.0
11	*due to the*	18.7	*as a result of*	6.0
12	*in the case*	17.9	*at the end of*	6.0
13	*on the other*	17.7	*is shown in fig*	5.8
14	*there is no*	16.7	*in the presence of*	5.6

Table 6. (*continued*)

Rank	3-gram	Norm freq*	4-gram	Norm freq*
15	*based on the*	16.5	*the fact that the*	5.6
16	*that it is*	16.4	*in the context of*	5.5
17	*part of the*	16.3	*the end of the*	5.4
18	*with respect to*	16.1	*as well as the*	5.1
19	*it is not*	15.6	*the results of the*	5.1
20	*a number of*	15.5	*can be used to*	4.5

* In order to facilitate comparison, all n-gram counts have been normalized to counts per 100,000 words.

Tables 4 and 5 display the 20 most common 3-grams and 4-grams in sets of MICUSP papers from different student levels. Items which appear in all four top-20 lists are set in small capitals. The first thing we observe is a considerable overlap of the items in all columns in Table 4. Eight out of 20 3-grams appear in the top lists of all four subcorpora (G0 to G3), however in different positions in the ranked lists. These items are: *in order to*, *as well as*, *the united states, one of the*, *due to the*, *in terms of*, *the use of* and *there is no*. Apart from *in order to* which is considerably more frequent in G0 papers than in G1/G2/G3, normalized frequencies of the shared items are in a comparable range for all four subsets. All eight items are commonly used by students at all MICUSP levels, mainly to introduce evaluation or to structure the discourse. There is additional overlap between levels G0 and G1 (*as a result*), between levels G0, G1 and G2 (*the number of*, *the fact that*, *there is a*, *be able to, part of the*), between levels G1, G2 and G3 (*based on the*, *it is not*), between levels G1 and G3 (*the role of*, *the relationship between*), and between levels G2 and G3 (*on the other*, *the other hand*). Only a few 3-grams are not shared by two or more lists but just occur among the top-20 3-grams in either G0, G1, G2 or G3 writing. *To be a*, *it would be* and the quantifiers *all of the* and *some of the* only occur in the final year undergraduate top-20 list. *It would be* here is generally used to introduce evaluation as in *it would be difficult to argue that the stories did not have an influence on the other* (ENG.G0.46.1) or to make a suggestion, as in *It would be interesting to include the size of the predator as one experimental factor* (BIO.G0.03.2).

The only 3-gram in the first year graduate student list that is not shared is *the importance of* which is particularly common in reports. Specific to the second year graduate level top-20 list are the items *more likely to*, *the development of* and *the effects of*, all three of which are common in the discussion of empirical studies or survey results (often in Economics and Psychology). *Of the state*, *the idea of*, *in other words* and *the case of* only appear among the top 20 items in the third year graduate 3-gram list and hence seem to be more prominent in the phraseological

repertoire of the most advanced students in our corpus. This does of course not mean that these list-specific 3-grams are not used at all by students at the other levels. It only means that items such as *the case of* or *the effect of* occupy a lower rank than 20 in the G0, G1 and G2 3-gram lists. *The effect of*, for example, ranks 117th in G0, 57th in G1, and 24th in the G2 list, so it appears that this particular item gains importance as students move from lower to higher levels, possibly because they more frequently discuss empirical studies and talk about causes and effects of factors and procedures at (advanced) graduate level.

The 3-gram *the effect of* does not appear in the Hyland frequency list in Table 6 but it only missed the top-20 by one rank. Not only is *the effect of* the 21st most frequent 3-gram in Hyland, the related cluster *the effects of* immediately follows it in position 22. A comparison of the Hyland top-20 3-grams and the MICUSP-based lists in Table 4 shows more similarities than differences between advanced student and expert academic writing when it comes to the use of phraseological items. Seven of the eight 3-grams that were found to be shared across the four MICUSP lists also made it into the Hyland top-20. There is further overlap between Hyland and the G0 list (five additional items: *the number of*, *the fact that*, *there is a*, *that it is*, *part of the*), between Hyland and G1 (six additional items: *the number of*, *the fact that*, *there is a*, *based on the*, *part of the*, *it is not*), between Hyland and G2 (seven additional items: *the number of*, *the fact that*, *there is a*, *on the other*, *based on the*, *part of the*, *it is not*), and between Hyland and G3 (five additional items: *the case of*, *on the other*, *based on the*, *that it is*, *it is not*). This means that between 12 and 14 top-20 3-grams are shared among Hyland and the four MICUSP subsets which implies that MICUSP students at all levels already use a large number of expert n-grams in their academic papers.

A similar picture emerges when we consider the top-20 lists of 4-grams in Table 5 and the Hyland 4-grams in Table 6. We observe overlaps of 9 to 11 items between the 20 most frequent Hyland 4-grams and the G0 to G3 lists. Among these shared items are: *in the case of*, *on the other hand*, *at the same time*, and *as a result of*. Again, our advanced student writers show more similarities than differences to expert writers in terms of their most commonly used phraseological items. Normalized frequencies occupy similar ranges for all corpora, too. If we only compare the four MICUSP datasets against each other, we observe an overlap of seven items among all lists (*in the united states*, *on the other hand*, *it is important to*, *as a result of*, *at the same time*, *in the case of*, *at the end of*), additional overlap between levels G0, G1 and G2 (*as well as the*, *the end of the*, *the university of michigan*, *at the university of*, *the fact that the*), and sets of items that occur in only one of the lists. This set of unshared items is particularly large for the third year graduate student list (level G3). Nine out of the top-20 4-grams in this subcorpus do not appear in any of the other three top-20 lists. These are: *of the united states*, *i would like to*, *for the*

purpose of, *as a function of*, *on the one hand*, *with respect to the*, *in addition to the*, *in the process of*, and *in a way that*.

Specific to the second year graduate list are the items *more likely to be*, *is more likely to* and *were more likely to*. The presence of these items in the G2 and G3 top-20 lists could indicate that the second and third year graduate student writers who contributed to MICUSP are more aware of a wider range of discourse structuring devices and especially evaluative phrases than their lower-level peers, as the following example from a third-year graduate paper in Economics illustrates: *This could be because answers of zero to 100 are more likely to reflect confusion than information about held beliefs* (ECO.G3.02.1). The one 4-gram that only occurs in the level G1 top-20 list is *the nature of the*, commonly used in reports as a qualifying expression, e.g. *depending on the nature of the research question* (NUR.G1.04.1). Finally, *the beginning of the*, *can be seen in*, *that there is a* and *to be able to* only occur in the final year undergraduate top-20 4-gram list. *The beginning of the* is topic-related in that it mainly refers to points in a literary work (in English essays) or in a lesson/school year (in Education response papers). *Can be seen in*, usually preceded by *as*, forming the 5-gram *as can be seen in*, frequently occurs in G0 research papers or proposals submitted by Mechanical Engineering students and is used to refer to figures or tables. Again, these level-specific 4-grams are not completely avoided by writers at the other levels. They only occur at lower (sometimes very low) ranks in the respective frequency-sorted lists.

While there are evidently massive similarities across levels with respect to the occurrence of high-frequency 3- and 4-grams, we observe some interesting differences when we put these n-grams back into context and look at their immediate collocates. To give a few examples, the ratio of R1 collocates (i.e. words in the first position to the right) of the 3-gram *in order to* increases steadily from level G0 (41.5%) to level G3 (71.8%). This means that, although absolute token numbers are lower, there is considerably more variation in terms of verb choice at higher (G2, G3) levels, and lower level MICUSP writers repeatedly use items from a smaller set of clusters, for example *in order to make* or *in order to be*. The same trend can be observed for other 3-grams that are shared across levels (e.g. *due to the, there is no*) and for common 4-grams, including *as a result of* and *in the case of*. While the most frequent collocates in R1 position for both 4-grams are quite similar at all four levels (mainly articles and personal pronouns), type-token ratios differ considerably and increase from 50.6% (G0) over 53.2% (G1) and 64.7% (G2) to 76.2% (G3) for *as a result of* collocates, and from 63.5% (G0) over 71.2% (G1) and 83.4% (G2) to 94.1% (G3) for the R1 collocates of *in the case of*.

3.2.2 *P-frames and their variants across MICUSP student levels*

In order to find out more about which phraseological items MICUSP writers at different levels use to structure their discourse, we will now turn from n-grams to phrase-frames and look at pattern variability. By highlighting which components of a string of words are fixed and which are flexible, p-frames summarize n-grams in a motivated way in that they tend to eliminate topic-specific items to a large extent while highlighting discourse items. Tables 7 and 8 present top-20 lists of high-frequency phrase-frames based on 3-grams (short "3-p-frames") and based on 4-grams (short "4-p-frames") extracted from the four level-specific MICUSP subcorpora. As in Tables 4 and 5, items which occur across all levels are set in small capitals. Only well-dispersed items with hits in at least five different MICUSP texts have been included in the lists. As in the n-gram analysis, p-frame lists derived from the Hyland corpus have also been considered and are referred to in the discussion of results below. The actual lists are not displayed here for reasons of space.

In Table 7, fifteen out of twenty 3-p-frames are shared across all level lists. These are *the * of*, *to * the*, *a * of*, *the * and*, *in * to*, *is * to*, *the * to*, *the * that*, *be * to*, *the * in*, *of * and*, *the * is*, *in * of*, *that * is*, and *it * be*. The same 15 items also appear in a top-20 3-p-frame list based on Hyland, so again there are strong similarities between apprentice and expert preferences for common academic phrases. Among the remaining items, we observe partial overlap, e.g. between levels G0, G1 and G2 (*to * a*; also among the top-20 in Hyland) and between levels G1, G2 and G3 (*the * between*; also among the top-20 in Hyland). There is only a small number of 3-p-frames which are not shared by two or more lists. *To * that*, with the most common variants *say*, *ensure* and *note*, and *the * the* (most common variants: *way*, *time*) only occur in the G0 top-20 list, whereas *are * to* (most common variants: *able*, *likely*) is specific to the list based on G1 papers. The p-frame *is * the*, with the most common variants *that, in* and *not*, is specific to the G2 list. The p-frames *to * that* (top variants: *say, believe, note*), *it * not* (top variants: *is, does, was*), *and * of* (top variants: *that, use, management*), *and of * in* (top variants: *women, power, interest*) occupy comparatively high positions in the G3 list, so they seem to be preferred by our more advanced writers. Still, these items do also occur in the lower level lists, only further down in terms of rank order.

While the top-20 3-p-frame lists are very similar for all student levels, we observe some potentially interesting lexical shifts from the less to the more advanced levels when we look at the most frequent p-frame variants. For the 3-p-frame *the * of*, for example, some top variants (*amount* and *importance*) are shared only by the lower levels G0 and G1, and others (*effect* and *idea*) by the upper levels G2 and G3. Yet other variants (*development* and *end*) are shared by levels G1 and G2. A number of variants are, however, shared across the four levels. For the 3-p-frame

Table 7. Top-20 3-p-frames across MICUSP student levels

Rank	Final year UG	Norm freq*	1st year graduate	Norm freq*	2nd year graduate	Norm freq*	3rd year graduate	Norm freq*
1	THE * OF	934.1	THE * OF	932.1	THE * OF	880.6	THE * OF	873.3
2	TO * THE	214.4	TO * THE	183.1	A * OF	143.5	A * OF	143.7
3	A * OF	171.0	A * OF	160.5	TO * THE	140.1	TO * THE	124.1
4	THE * AND	118.2	THE * AND	104.2	THE * AND	81.5	IS * TO	74.2
5	IN * TO	102.0	IS * TO	82.1	IN * TO	65.3	IN * TO	73.4
6	IS * TO	85.9	IN * TO	80.5	IS * TO	62.8	THE * AND	69.5
7	THE * TO	85.0	OF * AND	78.1	THE * THAT	62.6	THAT * IS	59.7
8	THE * THAT	83.0	BE * TO	67.5	THE * TO	62.2	THE * THAT	59.7
9	*to * a*	80.9	THE * THAT	66.2	*to * a*	54.3	OF * AND	56.0
10	BE * TO	63.3	*to * a*	64.6	BE * TO	53.2	BE * TO	54.3
11	THE * IN	62.2	THE * TO	62.5	THE * IS	52.3	THE * TO	47.3
12	OF * AND	60.2	*as * as*	52.5	*as * as*	51.1	*to * that*	45.7
13	*as * as*	59.6	THE * IS	51.7	THE * IN	48.9	IN * OF	39.5
14	THE * IS	56.0	IN * OF	51.0	IN * OF	46.6	*it * not*	38.1
15	*and * the*	47.6	THE * IN	49.1	OF * AND	43.7	THE * IS	36.1
16	IN * OF	46.7	*and * the*	46.7	IT * BE	39.2	*and * of*	35.9
17	THAT * IS	44.8	*it * be*	40.5	*is * the*	34.5	*of * in*	35.3
18	*to * that*	42.0	*are * to*	38.4	*and * the*	34.2	IT * BE	34.5
19	*the * the*	41.9	*the * between*	38.1	THAT * IS	33.1	*the * between*	34.5
20	IT * BE	40.1	THAT * IS	37.3	*the * between*	31.1	THE * IN	33.1

* In order to facilitate comparison, all n-gram counts have been normalized to counts per 100,000 words.

*the * of*, for instance, the words *number*, *use* and *effects* are among the most frequent variants in all variant lists. The most common variants of this p-frame in Hyland are *number*, *use*, *effects* (shared with all MICUSP levels), *effect* (shared with G2 and G3), and *presence*. Shared variants for the p-frame *a * of* are *result*, *variety* and *sense* (also among the top variants in Hyland). The few level-unique variants found for this frame (among the top-20 variants) point towards topic-related phrases, e.g. *theory* and *combination* (level G1), *discussion* and *model* (level G2).

Moving on to 4-p-frames (see Table 8), we again notice a considerable amount of overlap across ranked lists, with fourteen out of twenty items shared across all four levels (*the * of the*, *in the * of*, *it is * to*, *at the * of*, *it is * that*, *on the * of*, *the * of a*, *of the * of*, *to the * of*, *as a * of*, *in * of the*, *the * of this*, and *on the * hand*). Again, the same fourteen items are also among the top-20 4-p-frames in Hyland.

In addition, we find partial overlap between the G0, G1 and G2 lists (*as * as the, in * to the*) and between G0, G1 and G3 (*for the * of, and the * of*; also in the Hyland top-20). There are two list-specific items in the final-year undergraduate student 4-p-frame list, *with the * of* (top variants: *idea*, *use*, *help*), *of the * and* (top variants: *study*, *world*, *time*), and two in the first-year graduate student list, *is a * of* (top variants: *function*, *picture*), *will be * to* (top variants: *able*, *used*). Three 4-p-frames in the second-year graduate student top-20 list are not shared by any other list. These are *are * likely to* (top variants: *more*, *less*), *the * in which* (top variants: *ways*, *way*), and *the * and the* (top variants: *people*, *client*). Finally, four 4-p-frames are specific to the third-year graduate student list in Table 8: *is not * to* (top variants: *up*, *difficult*), *that the * of* (top variants: *use*, *content*), *the growing * of* (top variants: *trend*, *use*), and *of * of the* (top variants: *life*, *each*).

A closer look at some of the 4-p-frames that appear in all four top-20 lists displayed in Table 8 highlights both cross-level similarities and differences in terms of variant occurrence. The variable slot in *it is * to*, for example, is most commonly filled by the adjectives *important*, *difficult*, *possible*, *necessary*, *hard* and *easy*. This applies to all four student levels (and the Hyland corpus in which these adjectives, apart from *hard*, are among the top six variants of *it is * to*, together with *rational*). For this item, there are hardly any words in the level-based top-10 variant lists that are specific to only one list (one of the few exceptions is *crucial* which occupies rank eight in the G1 list). So the p-frame *it is * to* is not only very frequent across all four levels, it also contains largely the same variants throughout. The very similar structure *it is * that*, however, shows a somewhat different behavior. Only two of the top-10 variants at G0, G1, G2 and G3 levels (*clear* and *possible*) are shared among all lists. Some show partial overlap, and others are specific to one list only. Among these items at levels G0, G1 and G2 are the adjectives *obvious*, *crucial* and *surprising*, all of which do not occur at all in the G3 variant list for *it is * that*. Apparently, some of our student writers at lower levels use realizations of this p-frame to express stronger and more personal opinions than their higher level peers. Thompson's (2009) BAWE-based observations on the increasing use of introductory *it* patterns including *it is * to* and *it is * that* from lower to higher student levels could not be confirmed with reference to our data — perhaps because MICUSP writing comes from generally higher level students than the papers included in BAWE.

3.2.3 *Summary*

Unlike our first case study on attribution which showed considerable variation across MICUSP disciplines, the analysis of recurring phraseological items across levels did not uncover any great differences. Instead, our examination of n-grams and p-frames in final-year undergraduate and first- through third-year graduate

Table 8. Top-20 4-p-frames across MICUSP student levels

Rank	Final year UG	Norm freq*	1st year graduate	Norm freq*	2nd year graduate	Norm freq*	3rd year graduate	Norm freq*
1	THE * OF THE	178.4	THE * OF THE	152.1	THE * OF THE	126.8	THE * OF THE	105.6
2	IN THE * OF	55.4	IN THE * OF	66.4	IN THE * OF	45.3	IN THE * OF	58.3
3	IT IS * TO	27.4	AT THE * OF	21.5	THE * OF A	21.6	ON THE * HAND	24.1
4	AT THE * OF	26.2	ON THE * OF	20.2	AT THE * OF	19.4	IT IS * TO	23.0
5	IT IS * THAT	24.6	IT IS * THAT	19.3	OF THE * OF	18.9	THE * OF A	17.9
6	ON THE * OF	20.2	OF THE * OF	18.4	AS A * OF	16.2	AT THE * OF	17.6
7	THE * OF A	20.0	THE * OF A	17.9	TO THE * OF	15.5	AS A * OF	17.1
8	OF THE * OF	19.1	A * OF THE	17.6	ON THE * HAND	14.0	*for the * of*	15.4
9	TO THE * OF	17.5	AS A * OF	16.7	IT IS * TO	13.3	IT IS * THAT	14.6
10	AS A * OF	16.4	IT IS * TO	16.3	IT IS * THAT	11.9	THE * OF THIS	13.2
11	A * OF THE	16.1	TO THE * OF	15.2	IN * OF THE	11.3	OF THE * OF	12.9
12	*for the * of*	14.5	*and the * of*	14.3	ON THE * OF	11.0	TO THE * OF	12.6
13	*and the * of*	14.4	*on the * hand*	13.7	A * OF THE	10.6	ON THE * OF	12.3
14	*as * as the*	14.3	*for the * of*	12.9	*as * as the*	8.8	*and the * of*	10.1
15	*with the * of*	12.9	*as * as the*	10.1	*the * that the*	8.6	A * OF THE	9.8
16	IN * OF THE	12.7	IN * OF THE	10.0	*in * to the*	8.6	*is not * to*	9.2
17	THE * OF THIS	11.3	*in * to the*	9.6	*are * likely to*	8.1	*that the * of*	9.0
18	*of the * and*	11.1	THE * OF THIS	9.4	THE * OF THIS	8.1	IN * OF THE	8.4
19	*in * to the*	11.1	*is a * of*	8.2	*the * in which*	7.0	*the grow- ing * of*	7.8
20	ON THE * HAND	10.2	*will be * to*	7.9	*the * and the*	6.8	*of * of the*	7.0

* In order to facilitate comparison, all n-gram counts have been normalized to counts per 100,000 words.

writing highlighted consistency across the sets of upper-level student papers that are captured in MICUSP, as well as between advanced student writing and published academic writing (Hyland comparisons). The kind of variation that was observed did not relate so much to the frequency and selection of common items such as *in order to* or *as a result of* but more to their actual use in context where we found higher degrees of lexical variation immediately to the right of these n-grams at G2 and G3 than at G0 and G1 levels. On the whole, however, it seems that MICUSP students have picked up the phraseological items that are core in academic writing by the time they get to senior undergraduate level. This suggests that the MICUSP variable "student level" is less strong than "discipline" so that it appears safe in MICUSP-based cross-disciplinary studies to group papers from all four levels together.

Of course, phraseological items are only one of a range of lexical-grammatical phenomena that can be studied in advanced student writing at different ranks. Results from other MICUSP analyses indicate mild cross-level variation, for example in the use of attended and unattended *this*. Römer & Wulff (2010) have found that percentages of attended *this* slowly (but consistently) increase from G0 to G3. Also, some interesting cross-level differences have been observed for a selection of high-frequency sentence-initial *this*+verb clusters (e.g. *this means*, *this seems*), see Wulff et al. (2012), and for the distribution of scare-quotes (Aull & Barcy 2010, Avanessian & Swales 2010). However, further studies of other linguistic phenomena will be necessary to see how strongly the "increasing demands through rank" referred to by Swales (see quote in Section 2) actually influence the language used by upper-level student writers.

4. Conclusion

The two case studies reported here have illustrated how social variables can be approached in studies using MICUSP. We have focused on two different variables to account for variation in the corpus, one for each case study. In the case of attribution, we have looked at discipline, and in the case of phraseological units, we have looked at student level.

The use of attribution was investigated across ten different disciplines represented in the corpus. Considerable differences were found between those disciplines, both in the frequency of attribution and in the form it takes (whether integral or non-integral), showing that the MICUSP students have been very much socialized into their different disciplines, at least from the perspective of referring to other sources. When a subset of the results was compared to data on published academic writing (Hyland 1999), very similar patterns were found in four out of

six disciplines, indicating that the MICUSP student writers behave in ways similar to the experts.

We then examined the ranking and use of high-frequency phraseological items of different lengths in groups of MICUSP papers from four different student levels and observed that, with respect to this particular language phenomenon, there are more similarities than differences across level or rank. Students at lower and higher MICUSP levels show a similar understanding of common items such as *in order to*, *at the same time* or *it is important to,* which may suggest that these items are acquired at an earlier stage in their academic careers.

If we return to the Swales quote describing the creation of MICUSP as "building a picture of disciplinary variation (or not) and of increasing demands through rank (or not) as samples are collected and analysed", we can ask which picture emerges based on the results of the two case studies. The results show that the disciplinary variation is there in the case of attribution, but they show weak support for the notion of "increasing demands through rank" — examined here in terms of student level — in the case of phraseological units.

In order to determine whether these observations (discipline triggers linguistic variation in advanced student writing, student level only does to a limited extent) are specific to the language features our case studies are based on or more universally valid, we would have to carry out a larger-scale investigation and look more systematically at a wider range of phenomena, both across MICUSP disciplines and across MICUSP levels. Also beyond the scope of this article, but clearly very valuable, would be a more detailed qualitative analysis of the use of attribution units and phraseological items by different individual writers or groups of writers (e.g. second-year graduate students of Biology). It would be interesting to see to what extent certain uses of citation forms or n-grams are idiosyncratic rather than conventionalized.

In this article, we have considered only two MICUSP variables. With its detailed markup and annotation (see O'Donnell & Römer, forthcoming), MICUSP also enables users to investigate other variables such as nativeness, gender, and text type. It remains to be seen how strong nativeness and gender are in this particular context of writing. The non-native speakers of English represented in the corpus will naturally have a very high level of English, as they are students at a US university (for which there are relatively strict entrance requirements with respect to English proficiency) who manage to produce A-grade papers, so they may not differ too much from their native-speaker peers (see also Römer 2009a). Another interesting avenue for further research is in genre variation. The coding of the 829 corpus files for text type enables researchers to focus on this variable and study whether or how language phenomena vary across types. Especially the question

of how (dis)similar graduate student writing at this level is to published scholarly writing can be explored from a text type perspective.

Despite its recent release, MICUSP has already generated research covering several areas of study. One strand of research has examined the use of what are traditionally conceived of as function words, such as attended and unattended *this* (Wulff et al. 2012, and Römer & Wulff 2010), and anticipatory or introductory *it*. By means of analyzing MICUSP data, papers by advanced German learners of English, and published academic articles, Römer (2009b) discusses potential connections between different introductory *it* patterns and the development of native and non-native speakers' academic writing proficiency. MICUSP has also been used to study systematically, not words, but textual marking in the form of scare-quotes (Aull & Barcy 2010, Avanessian & Swales 2010). Furthermore, the corpus has been used as a source of solid empirical data for qualitative discourse analysis: specifically, to create a taxonomy of discourse functions of metadiscourse (Ädel 2010a; forthcoming). Not only linguistic forms, textual marking and discourse functions have been studied, but also the positional variation of n-grams and phrase-frames, providing an application of some of Hoey's (2005) lexical priming claims to student academic writing (O'Donnell & Römer, in preparation).

MICUSP will help ameliorate somewhat the lack of generally available corpora of academic writing and enable more systematic studies of student writing. We see MICUSP as a versatile tool for different types of comparative research, allowing a range of research questions — relating to not only writing across disciplines or student levels, but also, for example, native-speaker and advanced non-native-speaker writing. In making MICUSP a shared resource, we expect that our anticipated uses of the corpus will also be supplemented by other exciting lines of research, all of which will help illuminate aspects of advanced student academic writing.

Notes

1. Swales' idea was inspired by Cheryl Geisler's writings on the "Great Divide" between layperson and expert (e.g. Geisler 1994).

2. The project was launched in late 2004 by Rita Simpson-Vlach and John Swales. Annelie Ädel was project leader from 2005 to 2007. From 2007 to 2011, the project was managed by Ute Römer, supported by Matthew Brook O'Donnell as technical project manager (from 2008 to 2011).

3. The definitions used in the classification process can be found on the MICUSP project website at http://micusp.elicorpora.info/micusp-paper-classification.

4. See e.g. Swales (2004: 13) for recent work pointing to the importance of "divisional and disciplinary differences" in how genres are ranked.

5. Geological Sciences, one of the Biological and Health Sciences, was initially included but taken off the list since we did not receive any submissions from Geological Sciences students.

6. This is based on figures available on the University of Michigan Office of Budget and Planning website at http://sitemaker.umich.edu/obpinfo/enrollment_and_fte.

7. The MICUSP Simple interface can be accessed at http://search-micusp.elicorpora.info/.

8. The footnote is in superscript in the original and refers to "3 Howard, Donald R. Writers and pilgrims: medieval pilgrimage narratives and their posterity (Berkeley, University of California Press, 1980), 17".

9. The bracketed number refers to "[3] R. Unanyan, M. Fleishhauer, B. Shore, and K. Bergmann. Robust creation and phase-sensitive probing of superposition states via stimulated raman adiabatic passage (stirap) with degenerate dark states. Optics Communications, 155:144–154, October 1998".

10. For the sake of brevity, "History and Classical Studies" is abbreviated to "History" in the following.

11. One could see this as Psychology showing its allegiance to both the "soft" sciences in using integral forms and the "hard" sciences in using non-integral forms.

12. The last category is quite broad and heterogeneous because the MICUSP disciplines Industrial & Operations Engineering and Civil & Environmental Engineering were collapsed, as were the Hyland disciplines of Electronic Engineering and Mechanical Engineering.

13. Naturally, it is necessary to delve more deeply into the formal realizations of attribution than we have done here in order to really pin down potential differences. Further details concerning the forms of attribution (such as verb choice in integral types) are discussed in Ädel & Garretson (2006).

14. *KfNgram* treats all variations of an n-gram with a single variable slot in any position of the n-gram as p-frames (e.g. *BCD, A*CD, AB*D and ABC* for the 4-gram ABCD).

References

Ädel, A. 2010a. "'Just to give you kind of a map of where we are going': A taxonomy of metadiscourse in spoken and written academic English". *Nordic Journal of English Studies*, 9 (2), 69–97.

Ädel, A. 2010b. "Using corpora to teach academic writing: Challenges for the direct approach". In M. C. Campoy-Cubillo, B. Belles-Fortuño & L. Gea-Valor (Eds.), *Corpus-based Approaches to ELT*. London: Continuum, 39–55.

Ädel, A. Forthcoming. "'What I want you to remember is': Audience orientation in monologic academic genres". In L. Brems, L. Ghesquière & F. Van de Velde (Eds.), Special issue on intersubjectivity for *English Text Construction.*

Ädel, A. & Garretson, G. 2006. "Citation practices across the disciplines: The case of proficient student writing". In M. C. Pérez-Llantada Auría, R. Pló Alastrué & C. P. Neumann,

*Academic and Professional Communication in the 21*st *century: Genres, Rhetoric and the Construction of Disciplinary Knowledge*. Proceedings of the 5th International AELFE Conference, 271–280.

Alsop, S & Nesi, H. 2009. "Issues in the development of the British Academic Written English (BAWE) corpus". *Corpora*, 4 (1), 71–83.

Aull, L. & Barcy, K. 2010. "The 'good', the 'bad', and the snarky: Native and non-native student use of scare quotes in upper level academic writing". Paper presented at the 6th *Conference on Intercultural Rhetoric and Discourse, June 2010, Georgia State University, Atlanta, GA*.

Avanessian, N. & Swales, J. M. 2010. "Scare-quotes in MICUSP: Some preliminary observations". MICUSP Kibbitzer, available at http://micusp.elicorpora.info/micusp-kibbitzers (accessed December 2011).

Biber, D., Conrad, S. & Cortes, V. 2004. "*If you look at*: Lexical bundles in university teaching and textbooks". *Applied Linguistics*, 25 (3), 371–405.

Campbell, C. 1990. "Writing with others' words: Using background reading text in academic compositions". In B. Kroll (Ed.), *Second Language Writing. Research Insights for the Classroom*. Cambridge: Cambridge University Press, 211–230.

Dahl, T. 2004. "Textual metadiscourse in research articles: A marker of national culture or of academic discipline?". *Journal of Pragmatics*, 36, 1807–1825.

Ebeling, S & Heuboeck, A. 2007. "Encoding document information in a corpus of student writing: The experience of the British Academic Written English (BAWE)". *Corpora*, 2 (2), 241–256.

Fletcher, W. 2002–2007. *KfNgram*. Annapolis, MD: United States Naval Academy.

Geisler, C. 1994. "Literacy and expertise in the academy". *Language and Learning Across the Disciplines*, 1 (1), 35–56.

Granger, S. & Meunier, F. (Eds.) 2008. *Phraseology: An Interdisciplinary Perspective*. Amsterdam/Philadelphia: John Benjamins.

Hargens, L. L. 2000. "Using the literature: Reference networks, reference contexts, and the social structure of scholarship". *American Sociological Review*, 65 (6), 846–865.

Hoey, M. 2005. *Lexical Priming. A New Theory of Words and Language*. London: Routledge.

Hunston, S. & Francis, G. 2000. *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins.

Hyland, K. 1998. *Hedging in Scientific Research Articles*. Amsterdam/Philadelphia: John Benjamins.

Hyland, K. 1999. "Academic attribution: Citation and the construction of disciplinary knowledge". *Applied Linguistics*, 20 (3), 341–367.

Hyland, K. 2008. "Academic clusters: Text patterning in published and postgraduate writing". *International Journal of Applied Linguistics*, 18 (1), 41–62.

Krishnamurthy, R. & Kosem, I. 2007. "Issues in creating a corpus for EAP pedagogy and research". *Journal of English for Academic Purposes*, 6 (4), 356–373.

Michigan Corpus of Upper-level Student Papers. 2009. Ann Arbor, MI: The Regents of the University of Michigan.

Nesi, H., Sharpling, G. & Ganobcsik-Williams, L. 2004. "Student papers across the curriculum: Designing and developing a corpus of British student writing". *Computers and Composition*, 21, 439–450.

O'Donnell, M. B. & Römer, U. Forthcoming. "From student hard drive to web corpus (Part 2): The annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP)". *Corpora*.

O'Donnell, M. B. & Römer, U. In preparation. "Investigating the interaction between phraseological items and textual position".

Pecorari, D. 2006. "Visible and occluded citation features in postgraduate second-language writing". *English for Specific Purposes*, 25 (1), 4–29.

Römer, U. 2009a. "English in academia: Does nativeness matter?". *Anglistik: International Journal of English Studies*, 20 (2), 89–100.

Römer, U. 2009b. "The inseparability of lexis and grammar: Corpus linguistic perspectives". *Annual Review of Cognitive Linguistics*, 7, 140–162.

Römer, U. 2010. "Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews". *English Text Construction*, 3 (1), 95–119.

Römer, U. & O'Donnell, M. B. 2011. "From student hard drive to web corpus (Part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP)". *Corpora*, 6 (2), 159–177.

Römer, U. & O'Donnell, M. B. In preparation. *MICUSP: A Corpus Resource for Exploring Proficient Student Writing across Disciplines* (Book and CD-ROM; provisional title). Amsterdam/Philadelphia: John Benjamins.

Römer, U. & Schulze, R. (Eds.) 2009. *Exploring the Lexis-Grammar Interface*. Amsterdam/Philadelphia: John Benjamins.

Römer, U. & Wulff, S. 2010: Online. "Applying corpus methods to written academic texts: Explorations of MICUSP". *Journal of Writing Research*, 2 (2), 99–127. Available at: http://www.jowr.org/articles/vol2_2/JoWR_2010_vol2_nr2_Roemer_Wulff.pdf (accessed December 2011).

Simpson, R., Briggs, S., Ovens, J. & Swales, J. 1999. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Sinclair, J. M. 1987. "The nature of the evidence". In J. M. Sinclair (Ed.), *Looking up: An Account of the COBUILD Project in Lexical Computing*. London: HarperCollins, 150–159.

Sinclair, J. M. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, J. M. 1996. "The search for units of meaning". *Textus*, IX (1), 75–106.

Sinclair, J. M. 2004. *Trust the Text. Language, Corpus and Discourse*. London: Routledge.

Sinclair, J. M. 2008. "The phrase, the whole phrase, and nothing but the phrase". In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective*. Amsterdam/Philadelphia: John Benjamins, 407–410.

Stubbs, M. 2001. *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Swales, J. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, J. 2004. *Research Genres: Exploration and Applications*. Cambridge: Cambridge University Press.

Thompson, P. 2009. "Shared disciplinary norms and individual traits in the writing of British undergraduates". In E. Gotti (Ed.), *Commonality and Individuality in Academic Discourse*. Bern: Peter Lang, 53–82.

Wulff, S., Römer, U. & Swales J. M. 2012. "Attended/unattended *this* in academic writing: Quantitative and qualitative perspectives". *Corpus Linguistics and Linguistic Theory*, 8 (1), 129–157.

Authors' addresses

Annelie Ädel
Department of English
Stockholm University
S-106 91 Stockholm
Sweden

annelie.adel@english.su.se

Ute Römer
Department of Applied Linguistics and ESL
Georgia State University
34 Peachtree Street, Suite 1200
Atlanta, GA
United States of America

uroemer@gsu.edu